

# Gut community structure as a risk factor for infection in *Klebsiella pneumoniae*-colonized patients

Jay Vornhagen,<sup>1</sup> Krishna Rao,<sup>2</sup> Michael A. Bachman<sup>3,4</sup>

**AUTHOR AFFILIATIONS** See affiliation list on p. 16.

**ABSTRACT** The primary risk factor for infection with members of the *Klebsiella pneumoniae* species complex is prior gut colonization, and infection is often caused by the colonizing strain. Despite the importance of the gut as a reservoir for infectious *K. pneumoniae*, little is known about the association between the gut microbiome and infection. To explore this relationship, we undertook a case-control study comparing the gut community structure of *K. pneumoniae*-colonized intensive care and hematology/oncology patients. Cases were *K. pneumoniae*-colonized patients infected by their colonizing strain ( $N = 83$ ). Controls were *K. pneumoniae*-colonized patients who remained asymptomatic ( $N = 149$ ). First, we characterized the gut community structure of *K. pneumoniae*-colonized patients agnostic to case status. Next, we determined that gut community data is useful for classifying cases and controls using machine learning models and that the gut community structure differed between cases and controls. *K. pneumoniae* relative abundance, a known risk factor for infection, had the greatest feature importance, but other gut microbes were also informative. Finally, we show that integration of gut community structure with bacterial genotype data enhanced the ability of machine learning models to discriminate cases and controls. Interestingly, inclusion of patient clinical variables failed to improve the ability of machine learning models to discriminate cases and controls. This study demonstrates that including gut community data with *K. pneumoniae*-derived biomarkers improves our ability to classify infection in *K. pneumoniae*-colonized patients.

**IMPORTANCE** Colonization is generally the first step in pathogenesis for bacteria with pathogenic potential. This step provides a unique window for intervention since a given potential pathogen has yet to cause damage to its host. Moreover, intervention during the colonization stage may help alleviate the burden of therapy failure as antimicrobial resistance rises. Yet, to understand the therapeutic potential of interventions that target colonization, we must first understand the biology of colonization and if biomarkers at the colonization stage can be used to stratify infection risk. The bacterial genus *Klebsiella* includes many species with varying degrees of pathogenic potential. Members of the *K. pneumoniae* species complex have the highest pathogenic potential. Patients colonized in their gut by these bacteria are at higher risk of subsequent infection with their colonizing strain. However, we do not understand if other members of the gut microbiota can be used as a biomarker to predict infection risk. In this study, we show that the gut microbiota differs between colonized patients who develop an infection versus those who do not. Additionally, we show that integrating gut microbiota data with bacterial factors improves the ability to classify infections. Surprisingly, patient clinical factors were not useful for classifying infections alone or when added to microbiota-based models. This indicates that the bacterial genotype and the microbial community in which it exists may determine the progression to infection. As we continue to explore colonization as an intervention point to prevent infections in individuals colonized by

**Editor** Nicholas Chia, Argonne National Laboratory, Lemont, Illinois, USA

Address correspondence to Jay Vornhagen, jayvornh@iu.edu, or Michael A. Bachman, mikebach@med.umich.edu.

Krishna Rao is supported in part from an investigator-initiated grant from Merck & Co, Inc.; he has consulted for Seres Therapeutics, Inc., Rebiotix, Inc., and Summit Therapeutics, Inc. All other authors declare that they have no competing interests.

See the funding table on p. 16.

**Received** 10 June 2024

**Accepted** 11 June 2024

**Published** 8 July 2024

Copyright © 2024 Vornhagen et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

potential pathogens, we must develop effective means for predicting and stratifying infection risk.

**KEYWORDS** *Klebsiella*, microbiome

The gut is a vast ecosystem populated by trillions of bacteria, viruses, and microbial eukaryotes. The majority of these microbes have beneficial or neutral impacts on host health; however, some are potential pathogens. Under specific circumstances, some gut microbes can escape to distant body sites, leading to infection. One such group of pathogens is the *Klebsiella pneumoniae* species complex (referred to as “*K. pneumoniae*”). This complex contains several potentially pathogenic species of *Klebsiella*, including *K. pneumoniae*, *K. variicola*, *K. quasipneumoniae*, *K. quasivariicola* sp. nov., and *K. africana* [reviewed in reference (1)]. These bacteria are common causes of bacteremia, pneumonia, and urinary tract infection (UTI). The genome content of a given strain of *K. pneumoniae* determines its infectious potential, where the presence of virulence and fitness factors permits and enhances infectivity, and antimicrobial resistance genes complicate infection treatment (2). As of 2019, *K. pneumoniae* is the third leading global cause of death attributable to, or associated with, antimicrobial resistance (3). More research is necessary to understand *K. pneumoniae* pathogenesis. Such research may lead to improved diagnosis and treatment, and therein reduce the burden of *K. pneumoniae* disease.

*K. pneumoniae*-colonized patients are at increased risk for subsequent infection (4–6). Though few patient-centered studies determine the specific origin of infectious *K. pneumoniae*, those that have demonstrated that *K. pneumoniae*-colonized patients are infected with their colonizing strains in about ~80% of cases (4, 6, 7). Additionally, gut dominance by *K. pneumoniae* is a risk factor for infection in *K. pneumoniae*-colonized patients (8–10). The identification and interrogation of factors that permit, enhance, or restrict *K. pneumoniae* gut colonization are receiving increased attention due to the clear importance of the gut as a reservoir for infectious *K. pneumoniae*. Recent laboratory-based studies have identified novel gut fitness factors (11–14), microbes that enhance colonization resistance (15, 16), and gut community structures that are permissive or restrictive to colonization (11, 17). Despite increased interest, studies aiming to understand gut ecology in *K. pneumoniae*-colonized patients are comparatively sparse, limiting the translatability of laboratory-based findings to real-world settings.

Previously, we performed a cohort study of over 1,900 *K. pneumoniae*-colonized patients in the intensive care and hematology/oncology units (7). The goal of this study was to identify patient variables associated with infection. Additionally, two corresponding nested case-control studies were performed to assess the role of gut dominance in *K. pneumoniae* infection (8) and to rigorously identify infection-associated *K. pneumoniae* factors (18). Here, we aimed to leverage this case-control cohort of patients to understand the gut ecology of *K. pneumoniae*-colonized patients and determine if gut community structure can improve infection classification in *K. pneumoniae*-colonized patients. We found that community structure can improve infection classification. Machine learning models trained using gut community structure data in tandem with *K. pneumoniae* genotype can discriminate *K. pneumoniae*-colonized patients that proceed to infection from those that remain asymptotically colonized. Additionally, this suggests that there are biologically meaningful interactions between *K. pneumoniae* and the gut microbes that dictate the outcome of colonization.

## RESULTS

### Description of study population

Two hundred and thirty-eight patients were originally selected from a cohort of 1,978 *K. pneumoniae*-colonized intensive care and hematology/oncology patients (7) for a nested case-control study to assess the role of gut colonization density as a risk factor for *K. pneumoniae* infection (8). Cohort identification, enrollment, clinical data extraction,

chart review, case definitions, and case-control matching criteria are described in detail elsewhere (7, 8, 18). Briefly, residual rectal swabs originally collected for vancomycin-resistant *Enterococcus* screening upon admission to the intensive care units and the oncology wards from May 2017 to September 2018 at the University of Michigan hospitals in Ann Arbor, MI were used. Swabs were collected sequentially as they were being processed and without regard to any characteristics other than availability and yield. Patients were enrolled in our larger cohort study if *K. pneumoniae* was isolated from their rectal swab (7). Up to three isolates were archived from each swab along with all *K. pneumoniae* isolates from subsequent clinical cultures. Patient *K. pneumoniae* clinical cultures were evaluated as potential cases for alignment with clinical definitions for infection (7). Then, Sanger sequencing of the *wzi* locus was performed on the clinical and rectal isolates of confirmed infections to determine if the clinical isolate was concordant with the colonizing strain in the same patient (7, 8, 18). For the present study, we selected 232 patients (83 cases, 149 controls, Table 1) from the previous study based on inclusion in our previous comparative genomics study and available DNA extracted from the rectal swab most proximal to the infection (18). Cases were defined as colonized patients who met clinical criteria for concordant infection with a *K. pneumoniae* strain that was detectable in the gut prior to infection (7). The most common infection type was bacteremia, followed by UTI and respiratory infection (Table 1). Controls were defined as asymptotically colonized patients with a negative clinical culture collected of the same type as that of the matching case. Cases were matched with controls based on rectal swab collection date, age, and sex (18). Rectal swab DNA was extracted from concordant infections and matched controls (8). 16S rRNA gene sequencing was performed using the method described by Kozich et al. (19).

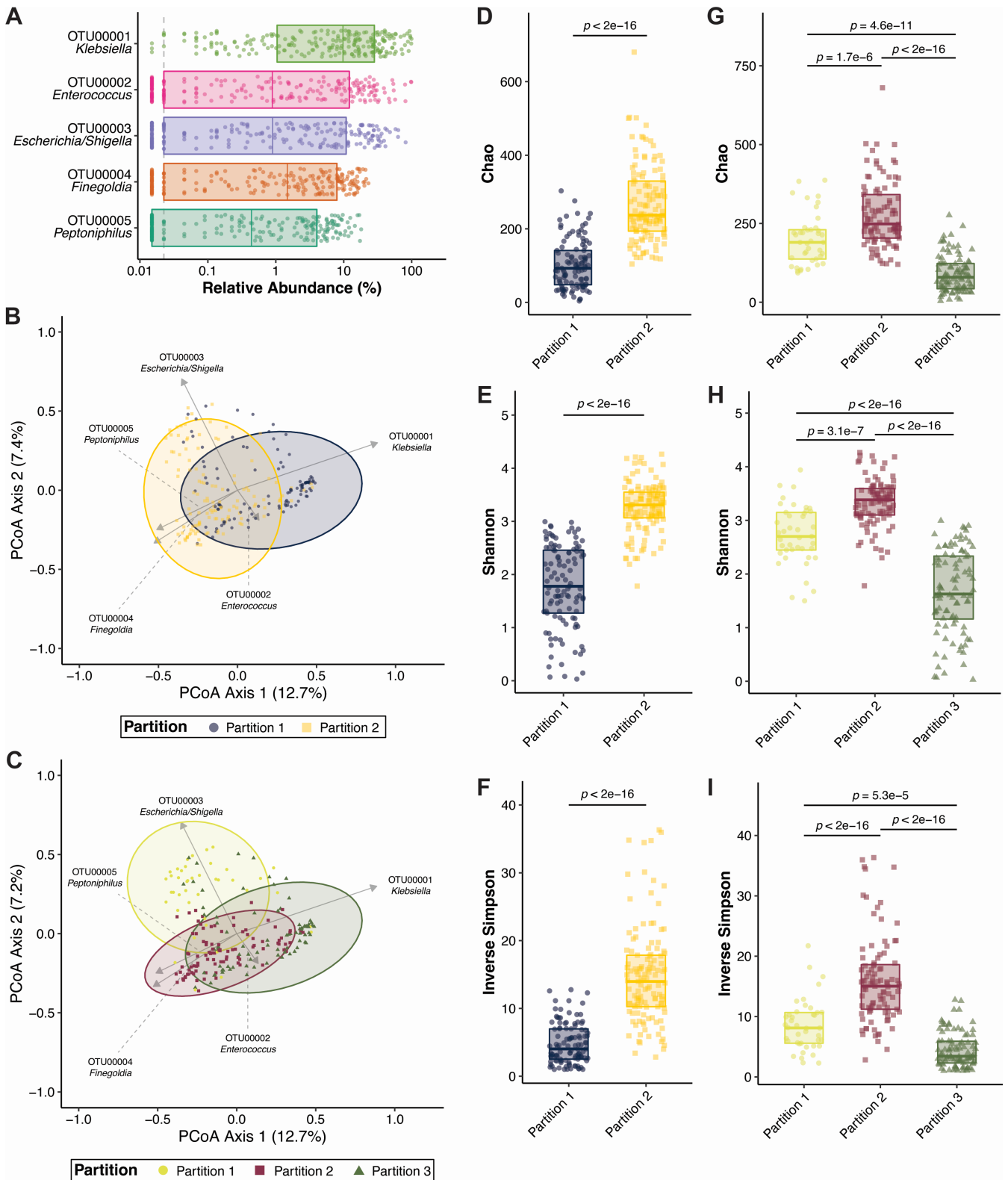
### The variation in gut community structure of *K. pneumoniae*-colonized patients is shaped by the relative abundances of *Escherichia*, *Klebsiella*, and anaerobes

First, we aimed to explore the gut community structure of *K. pneumoniae*-colonized patients agnostic of case status. *Klebsiella*, *Enterococcus*, *Escherichia/Shigella*, *Finnegoldia*, and *Peptoniphilus* were the dominant gut microbiota in this study population (Fig. 1A). Probabilistic modeling using Dirichlet multinomial mixtures (20) was used to determine if metacommunities exist in our study population. The optimal number of community

TABLE 1 Select patient demographics

| Variable               |                        | Case<br>(N = 83) | Control<br>(N = 149) | P value <sup>a</sup> |
|------------------------|------------------------|------------------|----------------------|----------------------|
| Age                    | Mean ± SD              | 60 ± 13          | 59 ± 12              | 0.556                |
| Sex                    | Male                   | 43 (51.8%)       | 76 (51.0%)           | 1.000                |
|                        | Female                 | 40 (48.2%)       | 71 (47.7%)           |                      |
|                        | Missing                | 0 (0%)           | 2 (1.3%)             |                      |
| Race/ethnicity         | African American       | 10 (12.0%)       | 19 (12.8%)           | 1.000                |
|                        | Asian                  | 2 (2.4%)         | 1 (0.7%)             | 0.292                |
|                        | Caucasian              | 70 (84.3%)       | 119 (79.9%)          | 0.482                |
|                        | Hispanic or Latino     | 1 (1.2%)         | 0 (0%)               | 0.358                |
|                        | Non-Hispanic or Latino | 68 (81.9%)       | 116 (77.8%)          | 0.503                |
|                        | Refused or missing     | 1 (1.2%)         | 3 (2.0%)             | 1.000                |
|                        | Other                  | 1 (1.2%)         | 5 (3.4%)             | 0.425                |
|                        | Hispanic or Latino     | 0 (0%)           | 1 (0.7%)             | 1.000                |
| Non-Hispanic or Latino | 1 (1.2%)               | 4 (2.7%)         | 0.657                |                      |
| Infection site         | Missing                | 0 (0%)           | 5 (3.4%)             | 0.163                |
|                        | Blood                  | 41 (49.4%)       |                      |                      |
|                        | Respiratory            | 19 (22.9%)       |                      |                      |
|                        | Urine                  | 23 (27.7%)       |                      |                      |

<sup>a</sup>Age: Student's *t* test; sex/race/ethnicity: Fisher's exact test.



**FIG 1** *Klebsiella pneumoniae* is the dominant gut microbe in *K. pneumoniae*-colonized patients. (A) Top five operational taxonomic units (OTUs) in *K. pneumoniae*-colonized patients ( $N = 232$ ). Principal coordinates analysis with overlaid biplots of OTUs of two- (B) and three-partition (C) community clustering using Dirichlet multinomial mixtures. Analysis of the Chao, Shannon, and Inverse Simpson alpha-diversity indices in two- (D-F) and three-partition community clustering (G-I, boxplot indicates median with interquartile range,  $P$  indicates Student's  $t$  test  $P$  value after Benjamini and Hochberg correction for multiple comparisons). For all panels, each data point indicates one patient.

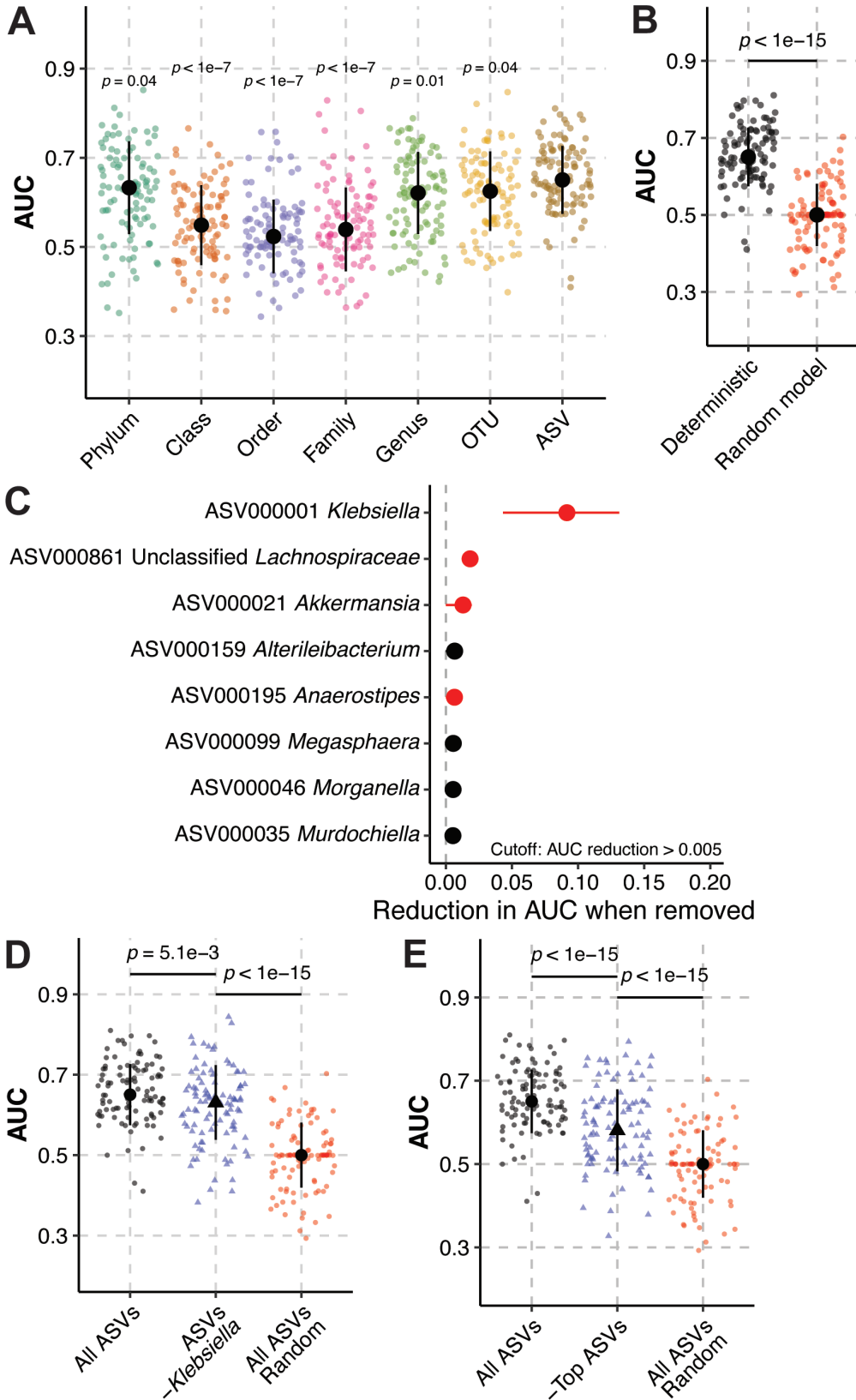
clusters was two (Laplace approximation = 194,340.97, Table S1), though one- and three-community clusters yielded similar fits (one-community Laplace approximation = 194,864.80, three-community Laplace approximation = 200,401.69, Table S1). Case status was not associated with metacommunity structure in either the two- [Partition 1 vs Partition 2: odds ratio (95% CI) = 1.35 (0.789–2.32)] or three-community models [Partition 1 vs Partition 2: odds ratio (95% CI) = 1.11 (0.505–2.46), Partition 1 vs Partition 3: odds ratio (95% CI) = 0.921 (0.414–2.05), Partition 2 vs Partition 3: odds ratio (95% CI) = 0.826 (0.46–1.48)]. Principal coordinates analysis revealed that OTU00001 *Klebsiella* and OTU00002 *Enterococcus* were strong components determining metacommunity structure in both two- (Partition 1, Fig. 1B) and three-partition communities (Partition 3, Fig. 1C), whereas other dominant gut microbiota influence different metacommunities (Fig. 1B and C). Alpha-diversity analysis of these metacommunities revealed that OTU00001 *Klebsiella* influenced partitions (Partition 1 and Partition 3 in two- and three-partition communities, respectively) were significantly less rich (Chao), even (Shannon) and diverse (Inverse Simpson) than other metacommunities (Fig. 1D through I). Interestingly, Partition 1 of the three-partition community clustering, which is heavily influenced by OTU00003 *Escherichia/Shigella* (Fig. 1C), was significantly less rich, even, and diverse than Partition 2 (Fig. 1G through I), which is influenced by OTU00004 *Fingoldia* and OTU00005 *Peptoniphilus* (Fig. 1C). Given that OTU00004 *Fingoldia* and OTU00005 *Peptoniphilus* are strict anaerobes and OTU00001 *Klebsiella*, OTU00002 *Enterococcus*, and OTU00003 *Escherichia/Shigella* are facultative anaerobes, it may be the case that alpha diversity is driven by the presence or absence of anaerobic bacteria in the gut in this patient population. Collectively, these data indicate that *K. pneumoniae* is the dominant gut microbe in this population of *K. pneumoniae*-colonized patients, and is associated with reduced richness, evenness, and diversity.

### Models using ASVs performed best at classifying cases and controls

Our next goal was to determine the ability of microbiota composition to discriminate cases from controls. To this end, we used supervised machine learning models to classify case status, using different taxonomic levels as input data. Prior to machine learning, the case and control data set was balanced ( $N = 83$ ) by randomly selecting 83 controls and all 83 cases for each model iteration. Due to their high interpretability compared to other methods, we chose to use regularized logistic regression. To ensure optimal model performance, training was iterated across several combinations of hyperparameters [as in reference (21)], wherein the hyperparameter combination that yielded peak training performance was used as the final model (Fig. S1). This process was repeated for phylum, class, order, family, genus, operational taxonomical units binned at 97% sequence similarity (OTU), and amplicon sequence variant (ASV) level data. ASVs provided the most robust discrimination of cases and controls, followed by OTU and phylum (Fig. 2A). Additionally, models using ASVs as their input variables were most likely to yield an AUC >0.5, indicating that the classification of cases and controls was better than random chance. Other model performance metrics yielded comparable results (Table 2). Deterministic elastic net models trained on ASV data significantly outperformed models where case and control status were randomized (Fig. 2B). As we observed optimal model performance with ASVs, we decided to use the taxonomic level for further study analyses.

### Classification does not solely rely on the top ASVs such as ASV000001 *Klebsiella*

Consistent with previous observations that gut dominance by *K. pneumoniae* is a risk factor for infection in colonized patients (8–10), ASV000001 *Klebsiella* was the most important feature in our regularized logistic regression models and was weighted toward cases (Fig. 2C). Interestingly, two other ASVs, ASV000021 *Akkermansia* and ASV000861 Unclassified *Lachnospiraceae*, were also highly important features weighted toward cases (Fig. 2C). This suggests that other members of the gut microbiota have discriminatory



**FIG 2** Amplicon sequence variants best discriminate cases and controls. Regularized logistic regression model performance, as measured by area under the receiver-operator characteristic curve (A, AUC) on 100 test data sets consisting of a random subset of samples (80%) to predict case status in *K. pneumoniae*-colonized patients using different taxonomical data inputs. (Continued on next page)

## FIG 2 (Continued)

All 83 cases and 83 randomly selected controls were used for each model. (B) Model performance, as measured by AUC, using ASVs as input data were compared to models where case and control status was randomized. (C) Top model features for regularized logistic regression models using amplicon sequence variants (ASVs) as input data "ASVs." Circles indicate mean feature importance and lines indicate interquartile range. Feature importance values in red and black indicate a regression weight that is weighted toward cases and controls, respectively. (D) Model performance, as measured by AUC, using ASVs as input data compared to a model where all *Klebsiella* ASVs were omitted "ASVs -*Klebsiella*". (E) Regularized logistic regression model performance on test data sets for 100 seeds predicting case status in *K. pneumoniae*-colonized patients using all ASVs (All ASVs) or excluding ASVs ASV000001, ASV000021, and ASV000816 ("-Top ASVs"). Black circles indicate median values, black lines indicate standard deviation, and *P* indicates Tukey multiple pairwise-comparison *P* value following one-way ANOVA compared to "ASV" (A, D, E) or Student's *t* test *P* value (B). For panels (A, B, C, and D), each data point indicates one test data set.

TABLE 2 Taxon level elastic net performance data<sup>b</sup>

|                    | Phylum                  | Class                   | Order                   | Family                  | Genus                   | OTU                     | ASV              |
|--------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|------------------|
| AUC <sup>a</sup>   | <b>0.62 (0.55–0.7)</b>  | <b>0.55 (0.48–0.61)</b> | <b>0.54 (0.5–0.58)</b>  | <b>0.55 (0.5–0.6)</b>   | <b>0.61 (0.55–0.69)</b> | <b>0.62 (0.55–0.68)</b> | 0.66 (0.61–0.71) |
| pRAUC <sup>a</sup> | 0.58 (0.52–0.64)        | <b>0.52 (0.48–0.58)</b> | <b>0.48 (0.47–0.54)</b> | <b>0.51 (0.47–0.55)</b> | 0.58 (0.51–0.63)        | 0.58 (0.52–0.62)        | 0.6 (0.55–0.64)  |
| Accuracy           | <b>0.59 (0.53–0.66)</b> | <b>0.52 (0.47–0.57)</b> | <b>0.52 (0.47–0.56)</b> | <b>0.54 (0.5–0.59)</b>  | <b>0.58 (0.53–0.66)</b> | <b>0.59 (0.53–0.66)</b> | 0.63 (0.59–0.69) |
| Sensitivity        | 0.6 (0.5–0.69)          | 0.54 (0.48–0.63)        | 0.5 (0.44–0.63)         | 0.52 (0.44–0.63)        | 0.57 (0.5–0.69)         | 0.62 (0.55–0.69)        | 0.55 (0.44–0.63) |
| Specificity        | <b>0.57 (0.5–0.69)</b>  | <b>0.51 (0.44–0.58)</b> | <b>0.54 (0.44–0.56)</b> | <b>0.57 (0.44–0.63)</b> | <b>0.59 (0.5–0.69)</b>  | <b>0.57 (0.5–0.69)</b>  | 0.71 (0.63–0.75) |
| PPV <sup>a</sup>   | <b>0.59 (0.53–0.65)</b> | <b>0.52 (0.47–0.57)</b> | <b>0.52 (0.47–0.56)</b> | <b>0.55 (0.5–0.6)</b>   | <b>0.59 (0.53–0.64)</b> | <b>0.59 (0.53–0.65)</b> | 0.66 (0.6–0.73)  |
| NPV <sup>a</sup>   | 0.59 (0.53–0.67)        | <b>0.52 (0.47–0.58)</b> | <b>0.52 (0.47–0.56)</b> | <b>0.54 (0.5–0.59)</b>  | 0.59 (0.53–0.64)        | 0.6 (0.53–0.67)         | 0.62 (0.57–0.67) |

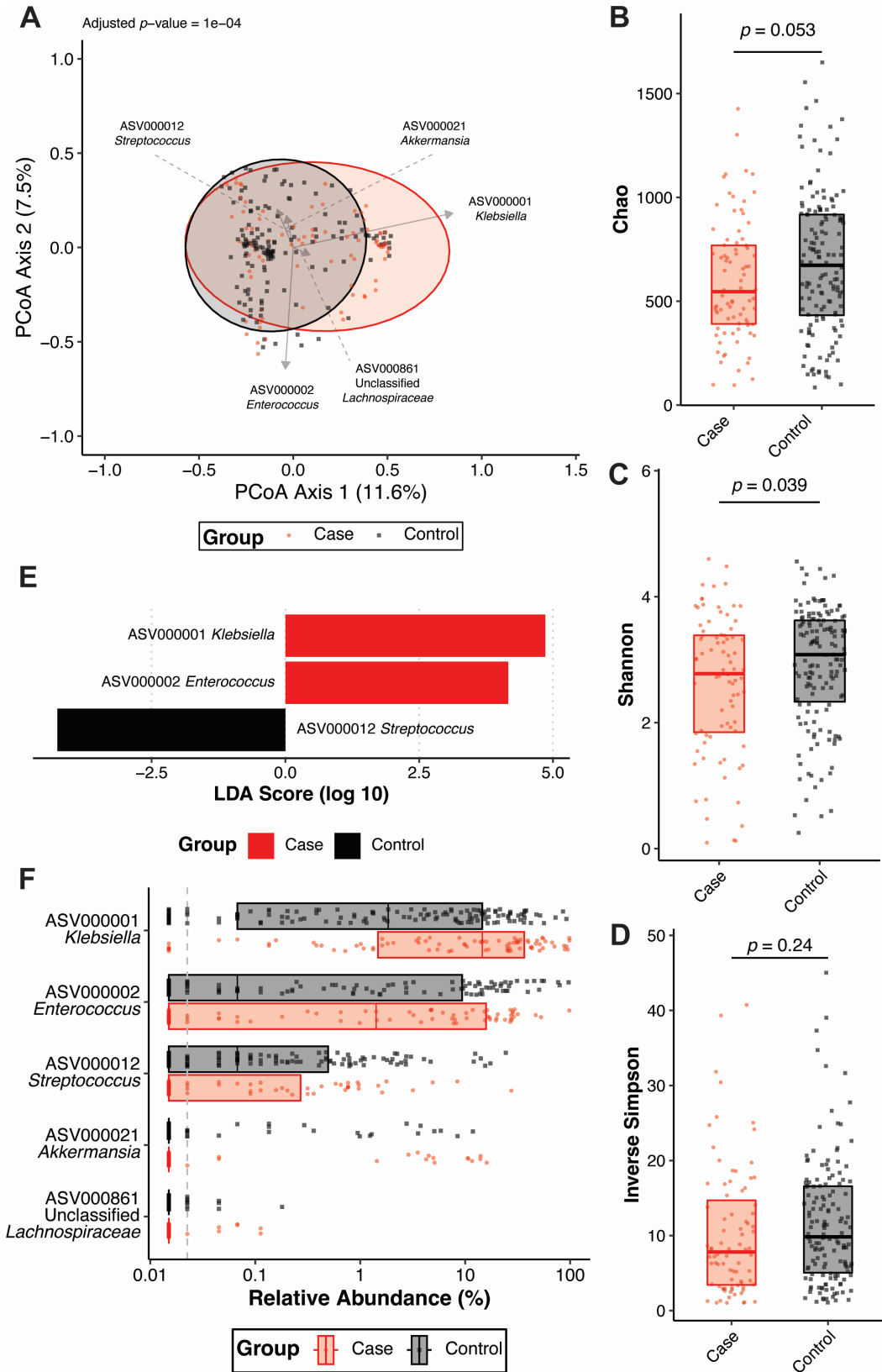
<sup>a</sup>AUC: area under the receiver-operating characteristic curve; PRAUC: area under the precision-recall curve; PPV: positive predictive value; NPV: negative predictive value.

<sup>b</sup>Median values and interquartile range are shown. Bold values are significantly different ( $P < 0.05$ ) from ASVs by Tukey multiple pairwise-comparison *P* value following one-way ANOVA. Tukey multiple pairwise-comparison *P* value following one-way ANOVA for all comparisons can be found in Table S5.

power for case status, rather than discriminatory power being limited to ASV000001 *Klebsiella*. Given the relatively high feature importance of these ASVs compared to other important features (Fig. 2C), we hypothesized that removal of the *Klebsiella* ASVs or ASVs with high feature importance may result in a model with no ability to classify cases and controls (AUC  $\leq 0.5$ ). Removal of all *Klebsiella* ASVs (Fig. 2D) or a combination of ASV000001, ASV000021, and ASV000861 (Fig. 2E) significantly reduced model performance; however, most models were still able to classify cases and controls better than chance (AUC  $> 0.5$ ) and outperformed a model where case status was randomly assigned. Similar results were observed upon removal of just ASV000001 *Klebsiella* (Fig. S2). This indicates that peak model performance does not rely solely on ASV000001, *Klebsiella*, or the three most important ASV features.

### Case and control gut community profiles differ

Given that cases and controls can be distinguished based on ASVs using machine learning, we next wanted to determine if the gut community profile of cases and controls differ. To this end, Yue and Clayton  $\theta$  dissimilarity index was calculated for each patient and used to assess the difference in beta-diversity between cases and controls. Visualization of distances using principal coordinates analysis revealed subtly different clustering of these groups (Fig. 3A). Though variance between the two groups was highly dimensional, as indicated by the low axis loadings (Fig. 3A), the gut microbiota of cases and controls was significantly different (adjusted *P*-value =  $1 \times 10^{-4}$ , Permutational multivariate analysis of variance, 1,000 permutations). Only community evenness (Shannon) was significantly different between cases and controls, though community richness and diversity displayed similar trends (Fig. 3B through D). Interestingly, the ASVs that were highly important for classifying cases and controls using machine learning models (Fig. 2B) partially differed from those enriched in either cases or controls. Linear discriminant analysis revealed that, as expected, ASV000001 *Klebsiella* was significantly enriched in cases, though unlike what was observed in the machine learning models, ASV000002 *Enterococcus* was also enriched in cases and ASV000012 *Streptococcus* was enriched in controls (Fig. 3E and F). Similar results were yielded using OTUs instead of



**FIG 3** Cases and controls have distinct gut community profiles based on ASVs (A) Principal coordinates analysis with overlaid biplots of specific ASVs. Permutational multivariate analysis of variance (PERMANOVA, 1,000 permutations) based on the Yue and Clayton  $\theta$  dissimilarity index was used to assess the difference in beta-diversity between cases ( $N = 83$ ) and (Continued on next page)

## FIG 3 (Continued)

controls ( $N = 149$ ). Analysis of the (B) Chao, (C) Shannon, and (D) Inverse Simpson alpha-diversity indices between cases ( $N = 83$ ) and controls ( $N = 149$ , boxplot indicates median with interquartile range,  $P$  indicates Student's  $t$  test  $P$  value). (E) Linear discriminant analysis (LDA) effect size was used to identify differentially abundant ( $P$  value  $< 0.05$ ) ASVs between cases ( $N = 83$ ) and controls ( $N = 149$ ). (F) Summary of relative abundances of ASVs that were differentially abundant (Fig. 3E) between cases ( $N = 83$ ) and controls ( $N = 149$ ) or highly important features for classification of cases and controls using regularized logistic regression shown in Fig. 2C (boxplot indicates median with interquartile range). For all panels, each data point indicates one patient.

ASVs to differentiate cases and controls (Fig. S3). Network analysis revealed that the gut community of controls was more connected than the gut community of cases (Fig. S4), suggesting a more stable gut community. Collectively, these data indicate that significant differences, not limited to *K. pneumoniae* relative abundance, exist between cases and controls that underpin the ability to discriminate these two groups based on gut community profile.

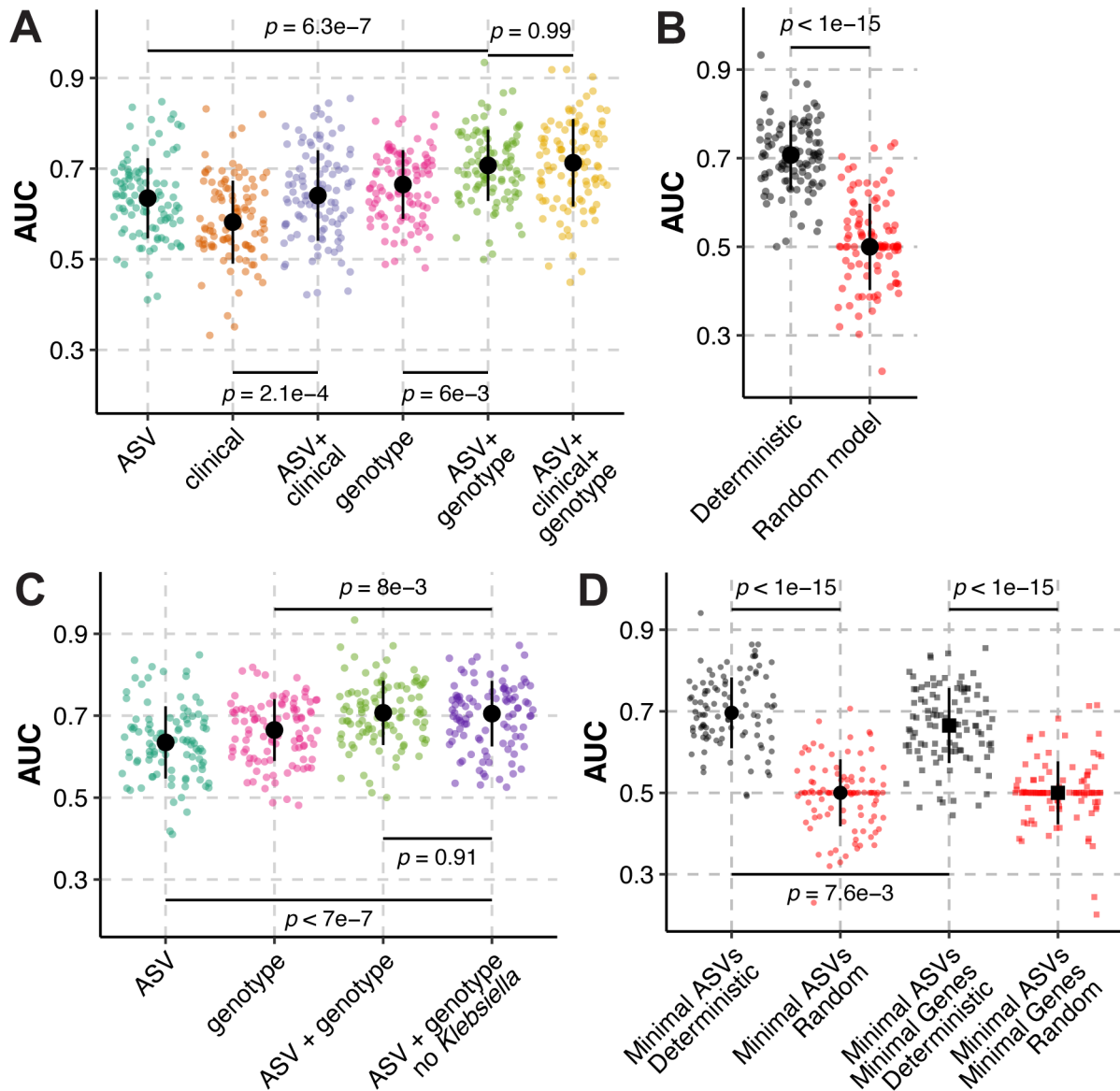
Previously, we detected the presence of multiple *K. pneumoniae* strains in colonized patients (18). A deeper exploration of ASVs revealed 30 ASVs that were classified as *Klebsiella*, and another 10,470 ASVs that were only classified to the level of Enterobacteriaceae. The majority (83.1%, 193/232) of patients had only one detectable *Klebsiella* ASV; however, 9.9% (23/232) of patients had multiple *Klebsiella* ASVs and 6.9% (16/232) had no *Klebsiella* ASVs (Fig. S5A) despite microbiological confirmation of *Klebsiella* colonization. The most abundant *Klebsiella* ASV was ASV000001, followed by ASV000019 (Fig. S5B). Importantly, only these two *Klebsiella* ASVs were included in the classification models in this study, as the rare *Klebsiella* ASVs were removed in the data preprocessing step prior to model training due to their near-zero variance between cases and controls. ASV000001 is 100% identical to the V4 region of all sequenced colonizing *K. pneumoniae* isolates in our original WGS study, except one, which had a contig break at 219 bp of the 253 bp V4 amplicon. Five samples in our data set with no detectable ASV000001 or ASV000019 had a detectable rare *Klebsiella* ASV ( $\leq 2$  reads). Given that all sequenced colonizing *K. pneumoniae* isolates had V4 amplicons identical to ASV000001, very low-abundance ASVs may be present due to sequencing errors; however, it is difficult to attribute rare ASVs to sequencing error with a high degree of confidence as only a single *K. pneumoniae* strain was sequenced from each colonized patient and co-colonization does occur. Interestingly, we detected ASV000019 *Klebsiella* in several controls, though no cases (Fig. S5B). Though ASVs do not provide high-confidence species-level resolution, it was notable that the ASV000019 16S rRNA gene sequence primarily aligned to members of the *K. oxytoca* complex (22), whereas the ASV000001 16S rRNA gene sequence primarily aligned to members of the *K. pneumoniae* complex (Table S2). Interestingly, ASV000001 is absent from patients colonized by ASV000019 (Fig. S5C). Collectively, these results suggest that species-level measurement of cocolonization may be possible through targeted genomic sequencing to understand colonization dynamics of high-abundance taxa, though it may not be possible at the strain-level

TABLE 3 ASV, clinical variable, and *K. pneumoniae* genotype elastic net performance data<sup>a</sup>

|                    | ASV              | Clinical                | Genotype                | ASV + clinical   | ASV + genotype          | ASV + clinical + genotype |
|--------------------|------------------|-------------------------|-------------------------|------------------|-------------------------|---------------------------|
| AUC <sup>a</sup>   | 0.66 (0.61–0.71) | <b>0.59 (0.53–0.66)</b> | 0.66 (0.61–0.72)        | 0.64 (0.58–0.71) | <b>0.71 (0.66–0.76)</b> | <b>0.71 (0.64–0.78)</b>   |
| prAUC <sup>a</sup> | 0.6 (0.55–0.64)  | 0.55 (0.5–0.6)          | <b>0.53 (0.48–0.57)</b> | 0.6 (0.54–0.65)  | <b>0.65 (0.61–0.7)</b>  | <b>0.66 (0.6–0.72)</b>    |
| Accuracy           | 0.63 (0.59–0.69) | <b>0.56 (0.5–0.6)</b>   | 0.63 (0.56–0.66)        | 0.6 (0.53–0.66)  | <b>0.65 (0.59–0.69)</b> | <b>0.65 (0.59–0.69)</b>   |
| Sensitivity        | 0.55 (0.44–0.63) | 0.52 (0.44–0.63)        | <b>0.47 (0.38–0.56)</b> | 0.57 (0.5–0.69)  | 0.57 (0.5–0.63)         | 0.58 (0.5–0.63)           |
| Specificity        | 0.71 (0.63–0.75) | 0.6 (0.5–0.69)          | <b>0.78 (0.69–0.88)</b> | 0.63 (0.56–0.7)  | <b>0.72 (0.63–0.81)</b> | <b>0.72 (0.63–0.81)</b>   |
| PPV <sup>a</sup>   | 0.66 (0.6–0.73)  | <b>0.57 (0.5–0.63)</b>  | <b>0.7 (0.62–0.75)</b>  | 0.61 (0.54–0.67) | <b>0.68 (0.62–0.73)</b> | <b>0.69 (0.6–0.75)</b>    |
| NPV <sup>a</sup>   | 0.62 (0.57–0.67) | <b>0.56 (0.5–0.6)</b>   | 0.6 (0.55–0.63)         | 0.6 (0.53–0.65)  | 0.63 (0.59–0.67)        | 0.63 (0.57–0.67)          |

<sup>a</sup>AUC: area under the receiver-operating characteristic curve; prAUC: area under the precision-recall curve; PPV: positive predictive value; NPV: negative predictive value.

<sup>b</sup>Median values and interquartile range are shown. Bold values are significantly different ( $P < 0.05$ ) from ASVs alone by Tukey multiple pairwise-comparison  $P$  value following one-way ANOVA. Tukey multiple pairwise-comparison  $P$  value following one-way ANOVA for all comparisons can be found in Table S6.



**FIG 4** Inclusion of ASVs enhances the ability to discriminate cases from controls. Regularized logistic regression model performance, as measured by area under the receiver-operator characteristic curve (A, AUC) on 100 test data sets consisting of a random subset of samples (80%) to classify case status in *K. pneumoniae*-colonized patients using combinations of clinical variables, *K. pneumoniae* genotype, and ASVs. (B) Model performance, as measured by AUC, using ASV and *K. pneumoniae* genotype as input data were compared to models where case and control status was randomized. (C) Model performance, as measured by AUC, using combinations of ASV and *K. pneumoniae* genotype as input data were compared to models where *Klebsiella* ASV000001 and ASV000019 were omitted “ASV + genotype no *Klebsiella*.” (D) Regularized logistic regression model performance, as measured by AUC, on 100 test data sets using ASV000001, ASV000002, ASV000012, ASV000021, ASV000861, and *K. pneumoniae* genotype (Minimal ASV) or using ASV000001, ASV000002, ASV000012, ASV000021, and ASV000861, and the five validated *K. pneumoniae* genes from reference (18) (Minimal ASV + Minimal Genes) as data inputs. Black circles indicate median values, black lines indicate standard deviation, and *P* indicates Tukey multiple pairwise-comparison *P* value following one-way ANOVA (A, C, D) or Student’s *t* test *P* value (B). For all panels, each data point indicates one patient.

with current approaches. More sophisticated sequencing techniques will need to be developed to assay colonization dynamics using discarded rectal swabs.

## Inclusion of gut microbiota data with *K. pneumoniae* genotype data enhances discrimination of cases and controls

Finally, we hypothesized that the inclusion of 16S rRNA gene sequencing data with clinical factors and *K. pneumoniae* genotype would enhance the ability of machine learning models to discriminate cases and controls. To test this hypothesis, we permuted ASVs with patient factors and *K. pneumoniae* genotype in our regularized logistic regression models. Eighty-four clinical factors, including several laboratory values, antibiotic exposure, and comorbidities, were included (Table S3) and the 27 infection-associated genes identified in our previous comparative genomics study were included as *K. pneumoniae* genotype (18). Clinical data were missing for two patients, so these patients were excluded from these analyses. Use of clinical factors as the sole input variables led to poor model performance (Table 3; Fig. 4A): 14/100 of regularized logistic regression models have an AUC  $\leq 0.5$ . Addition of ASVs to clinical factors enhanced median model performance (Fig. 4A). The lack of classifying ability of the clinical factors, especially antimicrobial exposure is somewhat surprising, as gut dominance is a known risk factor for infection (8–10), and disruption of the gut microbiota, such as what occurs with antibiotic exposure, leads to dominance in experimental gut colonization models (11, 13). Therefore, one may expect that antibiotic exposure would be an important feature for discriminating cases and controls in this study. Rather, exposure to most antibiotics was not among the most important features in regularized logistic regression models using clinical factors as the input variables (Fig. S6A) and the effects of antibiotic exposure on model performance was negligible (Fig. S6B). This included a variable for “high-risk” antibiotic exposure, which is a composite variable that includes  $\beta$ -lactam/ $\beta$ -lactamase inhibitor combinations, carbapenems, third- and fourth-generation cephalosporins, fluoroquinolones, clindamycin, and oral vancomycin based on their impact on indigenous gut microbiota (23). The only antibiotic present among the most important features was aminoglycoside exposure, and its effects on model performance were subtle (Fig. S6A). The importance of antibiotics was further reduced when ASVs were included (Figure S6C and D).

Use of *K. pneumoniae* genotype as the sole input variables led to a median model performance that was greater than that of clinical factors alone (Fig. 4A). This finding is expected, as 27 genes used as input variables are known to be associated with cases in our previous study (18), whereas most clinical factors were not associated with case status in our original cohort study (7). Interestingly, addition of ASVs to *K. pneumoniae* genotype enhanced model performance (Fig. 4A). Integration of all three data sets led to the highest median model performance, though performance was similar to models using only ASVs and *K. pneumoniae* genotype (Fig. 4A). Other model performance metrics yielded comparable results (Table 3), and use of OTUs with *K. pneumoniae* genotype subtly but significantly diminished performance compared to ASVs (Table S4). Importantly, our highest-performing model significantly outperformed a model where case and control status were randomized (Fig. 4B), and as was observed in Fig. 2D, omission of all ASV000001 and ASV000019 did not impact model performance (Fig. 4C).

Finally, we limited the input data to differential ASVs (Fig. 3F) and *K. pneumoniae* genotype. Model performance using limited ASVs was similar to that of inclusion of all ASVs (“Minimal ASVs,” Fig. 4D). Then, we limited *K. pneumoniae* genotype to the five genes we validated in our previous study using a geographically independent cohort of *K. pneumoniae*-colonized patients (18). Model performance using both the limited ASV and genotype sets also led to slightly reduced model performance compared to the model built on the complete data sets (“Minimal ASVs and Minimal Genes,” Fig. 4D). Both models performed better than a random model. In total, these data indicate that the inclusion of ASVs with *K. pneumoniae* genotype leads to peak model performance. This suggests that the gut community profile of patients colonized by *K. pneumoniae* can be combined with other infection-associated variables to discriminate, and potentially predict, infection in these patients with reasonable confidence.

## DISCUSSION

In this study, we have described the gut community of *K. pneumoniae*-colonized patients and demonstrated that the gut community differs between patients who remain asymptomatic (controls) and those who acquire a subsequent symptomatic infection with their colonizing strain (cases). In machine learning models built on this data, *K. pneumoniae* relative abundance had the greatest feature importance, though other gut microbes were also informative. Interestingly, clinical factors such as antibiotic exposure poorly discriminated cases and controls, whereas a combination of gut community data and *K. pneumoniae* genotype classified cases and controls more accurately than a random model (Fig. 4B), ASVs alone, or genotype alone (Fig. 4A). Collectively, this study demonstrates that the gut community of *K. pneumoniae*-colonized patients can be integrated with other biomarkers (patient factors or *K. pneumoniae* genotype) to assess infection risk. Moreover, these results suggest that there is a potentially important role for the gut microbiota community structure in determining the outcome of *K. pneumoniae* colonization.

Many microbiome studies classify individuals at risk for or experiencing disease as being in a state of dysbiosis; however, this imprecise term often lacks the context of the definition of a healthy microbiome. This is critical for establishing a causal link between the gut microbiome and disease, especially as the microbiome gradually shifts with age, environment, diet, healthcare exposure, and yet undiscovered variables [reviewed in reference (24)]. The goal of the present study is not to indicate that the gut microbiome of *K. pneumoniae*-colonized patients is in a state of health or dysbiosis. Rather, the goal is to identify biomarkers that classify infection in colonized patients. Ideally, the observations here will be tested experimentally to explore a causal role in disease. For example, *Akkermansia* (ASV00021) is currently being considered as a probiotic therapy due to its positive impacts on health (25–27). Yet, in this study, *Akkermansia* is important for model performance (Fig. 2C) but is relatively low abundance and not enriched in cases (Fig. 3C and D). This finding highlights differences between machine learning and classic linear discriminant analysis approaches for identifying sequences associated with specific communities. It may be the case that the ASVs identified through linear discriminant analysis have occult interactions with one another and/or other ASVs that explain the differential outcomes of these approaches. For example, recent experimental findings have determined antagonistic interactions exist between *K. pneumoniae* and *Escherichia coli* as a function of microbial diversity (28). Such interactions may be occurring in this study, as OTU00003 *Escherichia/Shigella* and OTU00001 *Klebsiella* were important and differential drivers of different gut community states (Fig. 1C). Similarly, laboratory experiments demonstrate that members of the *K. oxytoca* complex can reduce *K. pneumoniae* gut colonization (15). Here, we observed that the ASV that most likely represents the majority of the *K. pneumoniae* complex (ASV000001) is absent in patients colonized by the ASV that most likely represents the majority of the *K. oxytoca* complex (ASV000019, Fig. S5C). Despite a potential probiotic effect against *K. pneumoniae*, *K. oxytoca* is a pathogen that is often highly antimicrobial resistant [reviewed in reference (22)]. Therefore, while microbial competition with *K. pneumoniae* may explain this finding, characterizing *K. oxytoca* as a member of a healthy or dysbiotic gut microbiome remains in question. Further exploration of the gut community structures identified in this study is necessary to determine their importance in influencing infection risk in *K. pneumoniae*-colonized patients and therein define dysbiosis and its role in infection risk in this patient population.

The variables that are most important in classifying cases and controls likely differ between pathogens and patient populations. For example, clinical biomarkers do not appear to be critical for discriminating case status in this study (Fig. 4A; Fig. S5). This is in contrast to studies performed at the same clinical site leveraging electronic health records to stratify the risk of complicated *Clostridium (Clostridioides) difficile* infection (29), suggesting a disease-specific effect where the utility of these clinical data in making predictions varies across tasks. Notably, several of the most abundant OTUs described

here are consistent with previous cohorts at the same location, indicating some continuity of gut community structure within this geographical space (5). Other studies of *K. pneumoniae*-colonized individuals, including those colonized with multidrug-resistant (MDR) *K. pneumoniae* strains, report varying degrees of differing gut community structures, ranging from somewhat similar (30) to quite different (31, 32) than what is reported here based on most abundant OTUs. It is worth noting that it can be challenging to directly compare such studies due to differences in sample acquisition, data processing methods, and which data are reported. Similarly, the finding that ASVs yield the optimal taxonomical resolution for classifying case status (Fig. 2) is interesting. A recent machine learning study determined that OTUs were the optimal taxonomical level for predicting colorectal cancer (33). The preference for use of ASVs or OTUs in microbiome studies remains contested (34, 35); however, our study supports the premise that optimal taxonomical resolution is highly dependent on the patient population and outcomes of interest and does not necessarily favor OTUs or ASVs. Moreover, this study indicates that ASVs are not appropriately sensitive for strain-level data resolution for *K. pneumoniae*. This finding supports the growing body of research that culture-based approaches remain the gold standard for strain- and clone-level interrogation of gut microbial community structure (36). Ideally, clinical studies interrogating the role of the microbiome in disease would report both OTU and ASV data when using 16S rRNA gene sequencing. This would allow a better understanding of the role of taxonomical resolution in patient-based studies.

An important facet of this study population compared to other study populations is the diversity of colonizing *K. pneumoniae* strains. Often, studies aimed at describing the gut microbiota of *K. pneumoniae*-colonized patients capture patients colonized with highly clonal MDR *K. pneumoniae* strains (30, 37). In contrast, >100 unique sequence types of *K. pneumoniae* were identified in this study population, predominantly from non-MDR lineages (7). This may explain the lack of discriminatory power of antibiotic exposure for classifying cases and controls (Fig. S6). Antibiotic exposure may be a more discriminatory variable in cohorts of patients predominantly colonized by MDR strains. The attention given to MDR lineages is of course warranted; however, the majority of *K. pneumoniae* infections are caused by non-MDR lineages (38) and studies have demonstrated that the bulk of colonizing *K. pneumoniae* strains are diverse (39). Genetic differences in *K. pneumoniae* lineage may dictate interactions with gut microbiota that influence infection risk in patients colonized by MDR or hypervirulent *K. pneumoniae*. For example, we identified a *K. pneumoniae* factor canonically associated with hypervirulence, the *ter* operon, as a microbiome-dependent gut fitness factor (11). This locus was associated with infection in a hospital-wide patient cohort (40) but not in this cohort of intensive care and hematology/oncology patients (7). Alternatively, it may be that associations between *K. pneumoniae* and gut microbial community structure influence infection risk in a conserved manner. This would be ideal, as it would potentiate novel means for determining infection risk. This highlights the importance of studying all lineages with pathogenic potential to enable accurate risk assessment in colonized patients to reduce the burden of *K. pneumoniae* disease.

Though this study adds to our understanding of the gut microbiome of *K. pneumoniae*-colonized patients, it is not without its limitations. First, we used a case-control design for this study to carefully control for the influence of known and unknown patient factors. However, this study design leads to an overrepresentation of infection in the study population and the modeling metrics should be interpreted only in the context of this study, since in the general population we would expect a much lower infection risk, such as the 4.3% attack rate in our large cohort study from which this nested case-control study was derived (7). Ideally, future studies assessing the role of the microbiome as a risk factor for *K. pneumoniae* infection will accurately represent the true attack rate while capturing a large enough number of patients, both colonized by *K. pneumoniae* and not, to maintain suitable study power. Additionally, there may be uncaptured data, such as antibiotic use or post-discharge adverse healthcare events,

that occurred between the rectal swab collection and infection, as the duration between swab collection and infection ranged from 0 to 90 days (7). A comprehensive prospective cohort study that includes regular follow-up is necessary to test the ability of the models presented here to predict infection in colonized patients. Therefore, hypotheses generated in small- and medium-sized studies can be rigorously tested in a study population that reflects the general population. Second, this study is limited in its ability to make functional conclusions about the microbiome due to the use of 16S rRNA gene sequencing instead of metagenomics or other -omics approaches. Unfortunately, many -omics approaches remain cost-restrictive and lack easily testable hypotheses. This and similar studies will aid in the generation of hypotheses that can be tested using these approaches in the future. Third, some patients did not have detectable *Klebsiella* 16S DNA despite microbiological confirmation of colonization. It is possible that our sequencing effort was not deep enough, or DNA extraction was not efficient enough, to capture low-abundance *K. pneumoniae* events, whereas microbiological detection is more sensitive due to our use of selective and differential MacConkey agar. Finally, the use of machine learning models in this study is a useful means of determining the discriminatory ability of a large set of variables but is limited in its interpretability. Clinically actionable risk stratification models should be comprised of a small set of easily observable variables. In our previous studies, we developed practical tools for identifying biomarkers in *K. pneumoniae*-colonized patients including measurement of *K. pneumoniae* relative abundance and detection of infection-associated genes by PCR (7, 8). We hope that additional practical tools to assess the role of the microbiome in infection risk in *K. pneumoniae*-colonized patients will be developed and integrated with our previously developed tools.

The addition of this study to our collection of studies assessing patient factors, gut dominance, and *K. pneumoniae* genotype (7, 8, 18) represents one of the most comprehensive explorations of infection risk in a cohort of *K. pneumoniae*-colonized patients. Ultimately, this study provides a foundational framework for the development of integrated, actionable models for predicting and stratifying infection risk in *K. pneumoniae*-colonized patients.

## MATERIALS AND METHODS

### Study subject selection

Subjects in the present study were selected based on matching criteria, the availability of rectal swab DNA (8), and whole-genome sequencing data corresponding to the colonizing *K. pneumoniae*, *K. variicola*, or *K. quasipneumoniae* strain (18).

### 16S rRNA gene sequencing and data processing

DNA was previously extracted from patient rectal swabs (8) using the MagAttract PowerMicrobiome DNA/RNA Kit (Qiagen) and an epMotion 5075 liquid handling system. Standard PCRs used 1, 2, or 7  $\mu$ L of undiluted DNA and touchdown PCR used 7  $\mu$ L of undiluted DNA to amplify the V4 region of the 16S rRNA gene. Sequencing was performed as previously described (41). 16S rRNA gene sequences were processed with mothur (v.1.48.0) (19, 42). The sequencing error rate was assessed using a predefined mock community and estimated to be 0.033%. Sequences were aligned to the SILVA reference alignment, release 132 (43) and binned into OTUs using the OptiClust method (44) based on 97% sequence similarity or kept as unique sequences for ASVs. Taxonomic composition was assigned by classifying sequences within mothur using a modified version of the Ribosomal Database Project training set, version 18 (45, 46). Data processing was performed using the Great Lakes High-Performance Computing Cluster at the University of Michigan, Ann Arbor or the Carbonate large-memory computer cluster at Indiana University.

## Data analysis

Data analysis was carried out in RStudio 2021.09.0+351 “Ghost Orchid” Release for macOS or in R, v.4.2.0. R was used instead of RStudio when the analysis was being performed on The Great Lakes High-Performance Computing Cluster at the University of Michigan, Ann Arbor or the Carbonate large-memory computer cluster at Indiana University. For all analyses, sample read counts were rarefied to the lowest-abundance sample (4,438 reads). Alpha- and beta-diversity, principal coordinates analysis, and community typing were performed using *mothur*.  $\theta_{YC}$  was used as the distance metric for principal coordinates analysis. Differences in community structure were assessed by permutational multivariate analysis of variance (PERMANOVA, 1,000 permutations) from the *vegan* package, v.2.6-2 (47). Differences in alpha-diversity indices were assessed by Student’s *t* test using the *stats* package, v.3.6.2. Assessment of differentially enriched OTUs and ASVs was performed with linear discriminant analysis effect size analysis. Supervised machine learning was performed using *mikropml*, v.1.4.0 (48). First, continuous data were split into quartiles, then input data were preprocessed in *mikropml* using the default settings. All cases ( $N = 83$ ) and 83 controls were randomly selected for each model iteration. Supervised machine learning was performed using case status as the outcome. Input data were split 80:20 into train and test groups. An optimal model was trained using 100× 5-fold cross-validation and model performance was evaluated using the test data. For regularized logistic regression, hyperparameter selection was semi-automated. Each model was trained with alpha values ranging from 0 to 1, iterated in steps of 0.1, permuted with lambda values ranging from  $10^{-4}$  to  $10^1$ , iterated in steps of 3 between each log (e.g.,  $10^{-4}$ ,  $2.5 \times 10^{-4}$ ,  $5 \times 10^{-4}$ ,  $7.5 \times 10^{-4}$ ,  $10^{-3}$ ,  $2.5 \times 10^{-3}$ ... $10^1$ ). Hyperparameters that yielded the best performance were selected to evaluate model performance using the test data. This process was parallelized 100 times, using 100 different seeds to determine the train:test data split, and feature importance and weight was determined for all variables. Model performance metrics were defined as follows: Area under the receiver-operating characteristic curve (AUC), which is a calculation of the area under curve generated by plotting the true positive rate as a function of the false positive rate, wherein a value 0.5 is random classification of outcomes and 1.0 is perfect classification of outcomes; Area under the precision-recall curve (PRAUC), which is a calculation of the area under a curve generated by plotting precision (also known as positive predictive value) as a function of recall (also known as sensitivity), wherein a value 0.5 is random classification of outcomes and 1.0 is perfect classification of outcomes;

$$\text{Sensitivity} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}; \text{Specificity} = \frac{\text{True negatives}}{\text{True negative} + \text{False positives}}$$

$$\text{Positive predictive value (PPV)} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}; \text{Negative predictive value (NPV)}$$

$$= \frac{\text{True negatives}}{\text{True negatives} + \text{False negatives}}$$

Network analysis was performed using *NetCoMi* v.1.1.0 (49). Networks were constructed using the compositionally aware correlation estimators, *SparCC* (50), and networks were compared by permutation test with 100 permutations. For all analyses, a *P* value  $\leq 0.05$  after Benjamini-Hochberg adjustment was considered statistically significant. Data were visualized using *ggplot2*, v.4.1.2 (51).

## ACKNOWLEDGMENTS

The authors would like to thank the University of Michigan Microbiome Core for their assistance with 16S rRNA gene sequencing and Dr. Anna Seekatz for her valuable insight into all things gut microbiome. This research was supported in part through computational resources and services provided by Advanced Research Computing (ARC), a division of Information and Technology Services (ITS) at the University of Michigan, Ann Arbor. The authors acknowledge the Indiana University Pervasive Technology Institute for providing supercomputing and storage resources that have contributed to the research results reported within this paper.

This work was supported by funding from National Institutes of Health (<https://www.nih.gov/>) grants R01AI125307 to M.A.B. and R00 AI153483 to J.V. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conceptualization: J.V., K.R., M.A.B. Methodology: J.V., K.R. Investigation: J.V. Visualization: J.V. Funding acquisition: J.V., M.A.B. Project administration: M.A.B. Supervision: M.A.B. Writing – original draft: J.V. Writing – review and editing: J.V, K.R., M.A.B.

## AUTHOR AFFILIATIONS

<sup>1</sup>Department of Microbiology & Immunology, Indiana University School of Medicine, Indianapolis, Indiana, USA

<sup>2</sup>Department of Internal Medicine/Infectious Diseases Division, Michigan Medicine, University of Michigan, Ann Arbor, Michigan, USA

<sup>3</sup>Department of Pathology, Michigan Medicine, University of Michigan, Ann Arbor, Michigan, USA

<sup>4</sup>Department of Microbiology & Immunology, Michigan Medicine, University of Michigan, Ann Arbor, Michigan, USA

## AUTHOR ORCID*s*

Jay Vornhagen  <http://orcid.org/0000-0002-1685-302X>

Krishna Rao  <http://orcid.org/0000-0002-9213-7850>

Michael A. Bachman  <http://orcid.org/0000-0003-2507-6987>

## FUNDING

| Funder  | Grant(s) | Author(s)          |
|---|----------|--------------------|
| <a href="#">HHS   NIH   National Institute of Allergy and Infectious Diseases (NIAID)</a> | AI125307 | Michael A. Bachman |
| <a href="#">HHS   NIH   National Institute of Allergy and Infectious Diseases (NIAID)</a> | AI153483 | Jay Vornhagen      |

## DATA AVAILABILITY

The sequencing data generated in this study have been deposited in the Sequence Read Archive (SRA) database under accession [PRJNA789565](#). Deidentified human data are available under restricted access and can be obtained from M.A.B. within 1 year upon request, pending approval from the University of Michigan Institutional Review Board. All other source data and code are available at <https://github.com/jayvorn/Gut-community-structure-as-a-risk-factor-for-infection-in-Klebsiella-colonized-patients>.

## ETHICS APPROVAL

Patient enrollment and sample collection at the University of Michigan were approved by and performed per the Institutional Review Board (IRB) of the University of Michigan Medical School (Study number HUM00123033). This study was performed with a waiver of informed consent since the research involves no more than minimal risk to the subjects, could not practicably be carried out without the waiver, and uses discarded samples.

## ADDITIONAL FILES

The following material is available [online](#).

## Supplemental Material

Supplemental figures and tables (mSystems00786-24-s0001.docx). Fig. S1-S6 and Tables S1-S6.

## Open Peer Review

PEER REVIEW HISTORY (review-history.pdf). An accounting of the reviewer comments and feedback.

## REFERENCES

- Wyres KL, Lam MMC, Holt KE. 2020. Population genomics of *Klebsiella pneumoniae*. *Nat Rev Microbiol* 18:344–359. <https://doi.org/10.1038/s41579-019-0315-1>
- Martin RM, Bachman MA. 2018. Colonization, infection, and the accessory genome of *Klebsiella pneumoniae*. *Front Cell Infect Microbiol* 8:4. <https://doi.org/10.3389/fcimb.2018.00004>
- Antimicrobial Resistance Collaborators. 2022. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* 399:629–655. [https://doi.org/10.1016/S0140-6736\(21\)02724-0](https://doi.org/10.1016/S0140-6736(21)02724-0)
- Martin RM, Cao J, Brisse S, Passet V, Wu W, Zhao L, Malani PN, Rao K, Bachman MA. 2016. Molecular epidemiology of colonizing and infecting isolates of *Klebsiella pneumoniae*. *mSphere* 1:e00261-16. <https://doi.org/10.1128/mSphere.00261-16>
- Collingwood A, Blostein F, Seekatz AM, Wobus CE, Woods RJ, Foxman B, Bachman MA. 2020. Epidemiological and microbiome associations between *Klebsiella pneumoniae* and vancomycin-resistant *Enterococcus* colonization in intensive care unit patients. *Open Forum Infect Dis* 7:ofaa012. <https://doi.org/10.1093/ofid/ofaa012>
- Gorrie CL, Mirceta M, Wick RR, Edwards DJ, Thomson NR, Strugnell RA, Pratt NF, Garlick JS, Watson KM, Pilcher DV, McGloughlin SA, Spelman DW, Jenney AWJ, Holt KE. 2017. Gastrointestinal carriage is a major reservoir of *Klebsiella pneumoniae* infection in intensive care patients. *Clin Infect Dis* 65:208–215. <https://doi.org/10.1093/cid/cix270>
- Rao K, Patel A, Sun Y, Vornhagen J, Motyka J, Collingwood A, Teodorescu A, Baang JH, Zhao L, Kaye KS, Bachman MA. 2021. Risk factors for *Klebsiella* infections among hospitalized patients with preexisting colonization. *mSphere* 6:e0013221. <https://doi.org/10.1128/mSphere.00132-21>
- Sun Y, Patel A, SantaLucia J, Roberts E, Zhao L, Kaye K, Rao K, Bachman MA. 2021. Measurement of *Klebsiella* intestinal colonization density to assess infection risk. *mSphere* 6:e0050021. <https://doi.org/10.1128/mSphere.00500-21>
- Pérez-Nadales E, M Natera A, Recio-Rufián M, Guzmán-Puche J, Marín-Sanz JA, Martín-Pérez C, Cano Á, Castón JJ, Elías-López C, Machuca I, Gutiérrez-Gutiérrez B, Martínez-Martínez L, Torre-Cisneros J. 2022. Prognostic significance of the relative load of KPC-producing *Klebsiella pneumoniae* within the intestinal microbiota in a prospective cohort of colonized patients. *Microbiol Spectr* 10:e0272821. <https://doi.org/10.1128/spectrum.02728-21>
- Shimasaki T, Seekatz A, Bassis C, Rhee Y, Yelin RD, Fogg L, Dangana T, Cisneros EC, Weinstein RA, Okamoto K, Lolans K, Schoeny M, Lin MY, Moore NM, Young VB, Hayden MK. 2019. Increased relative abundance of *Klebsiella pneumoniae* Carbaapenemase-producing *Klebsiella pneumoniae* within the gut microbiota is associated with risk of bloodstream infection in long-term acute care hospital patients. *Clin Infect Dis* 68:2053–2059. <https://doi.org/10.1093/cid/ciy796>
- Vornhagen J, Bassis CM, Ramakrishnan S, Hein R, Mason S, Bergman Y, Sunshine N, Fan Y, Holmes CL, Timp W, Schatz MC, Young VB, Simner PJ, Bachman MA. 2021. A plasmid locus associated with *Klebsiella* clinical infections encodes a microbiome-dependent gut fitness factor. *PLoS Pathog* 17:e1009537. <https://doi.org/10.1371/journal.ppat.1009537>
- Hudson AW, Barnes AJ, Bray AS, Ornelles DA, Zafar MA. 2022. *Klebsiella pneumoniae* l-fucose metabolism promotes gastrointestinal colonization and modulates its virulence determinants. *Infect Immun* 90:e0020622. <https://doi.org/10.1128/iai.00206-22>
- Young TM, Bray AS, Nagpal RK, Caudell DL, Yadav H, Zafar MA. 2020. Animal model to study *Klebsiella pneumoniae* gastrointestinal colonization and host-to-host transmission. *Infect Immun* 88:e00071-20. <https://doi.org/10.1128/IAI.00071-20>
- Vornhagen J, Sun Y, Breen P, Forsyth V, Zhao L, Mobley HLT, Bachman MA. 2019. The *Klebsiella pneumoniae* citrate synthase gene, *gltA*, influences site specific fitness during infection. *PLoS Pathog* 15:e1008010. <https://doi.org/10.1371/journal.ppat.1008010>
- Osbelt L, Wende M, Almási É, Derksen E, Muthukumarasamy U, Lesker TR, Galvez EJC, Pils MC, Schalk E, Chhatwal P, Färber J, Neumann-Schaal M, Fischer T, Schlüter D, Strowig T. 2021. *Klebsiella oxytoca* causes colonization resistance against multidrug-resistant *K. pneumoniae* in the gut via cooperative carbohydrate competition. *Cell Host Microbe* 29:1663–1679. <https://doi.org/10.1016/j.chom.2021.09.003>
- Oliveira RA, Ng KM, Correia MB, Cabral V, Shi H, Sonnenburg JL, Huang KC, Xavier KB. 2020. *Klebsiella michiganensis* transmission enhances resistance to *Enterobacteriaceae* gut invasion by nutrition competition. *Nat Microbiol* 5:630–641. <https://doi.org/10.1038/s41564-019-0658-4>
- Sequeira RP, McDonald JAK, Marchesi JR, Clarke TB. 2020. Commensal Bacteroidetes protect against *Klebsiella pneumoniae* colonization and transmission through IL-36 signalling. *Nat Microbiol* 5:304–313. <https://doi.org/10.1038/s41564-019-0640-1>
- Vornhagen J, Roberts EK, Unverdorben L, Mason S, Patel A, Crawford R, Holmes CL, Sun Y, Teodorescu A, Snitkin ES, Zhao L, Simner PJ, Tamma PD, Rao K, Kaye KS, Bachman MA. 2022. Combined comparative genomics and clinical modeling reveals plasmid-encoded genes are independently associated with *Klebsiella* infection. *Nat Commun* 13:4459. <https://doi.org/10.1038/s41467-022-31990-1>
- Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl Environ Microbiol* 79:5112–5120. <https://doi.org/10.1128/AEM.01043-13>
- Holmes I, Harris K, Quince C. 2012. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS One* 7:e30126. <https://doi.org/10.1371/journal.pone.0030126>
- Topçuoğlu BD, Lesniak NA, Ruffin MT, Wiens J, Schloss PD. 2020. A framework for effective application of machine learning to microbiome-based classification problems. *mBio* 11:e00434-20. <https://doi.org/10.1128/mBio.00434-20>
- Yang J, Long H, Hu Y, Feng Y, McNally A, Zong Z. 2022. *Klebsiella oxytoca* complex: update on taxonomy, antimicrobial resistance, and virulence. *Clin Microbiol Rev* 35:e0000621. <https://doi.org/10.1128/CMR.00006-21>
- Baggs J, Jernigan JA, Halpin AL, Epstein L, Hatfield KM, McDonald LC. 2018. Risk of subsequent sepsis within 90 days after a hospital stay by type of antibiotic exposure. *Clin Infect Dis* 66:1004–1012. <https://doi.org/10.1093/cid/cix947>
- Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. 2018. Current understanding of the human microbiome. *Nat Med* 24:392–400. <https://doi.org/10.1038/nm.4517>
- Plovier H, Everard A, Druart C, Depommier C, Van Hul M, Geurts L, Chilloux J, Ottman N, Duparc T, Lichtenstein L, Myrdakis A, Delzenne NM, Klievink J, Bhattarjee A, van der Ark KCH, Aalvink S, Martinez LO, Dumas M-E, Maier D, Loumaye A, Hermans MP, Thissen J-P, Belzer C, de Vos WM, Cani PD. 2017. A purified membrane protein from *Akkermansia muciniphila* or the pasteurized bacterium improves metabolism in obese and diabetic mice. *Nat Med* 23:107–113. <https://doi.org/10.1038/nm.4236>
- Dao MC, Everard A, Aron-Wisniewsky J, Sokolovska N, Prifti E, Verger EO, Kayser BD, Levenez F, Chilloux J, Hoyle L, Dumas M-E, Rizkalla SW, Doré

- J, Cani PD, Clément K, MICRO-Obes Consortium. 2016. *Akkermansia muciniphila* and improved metabolic health during a dietary intervention in obesity: relationship with gut microbiome richness and ecology. *Gut* 65:426–436. <https://doi.org/10.1136/gutjnl-2014-308778>
27. Bae M, Cassilly CD, Liu X, Park S-M, Tusi BK, Chen X, Kwon J, Filipčič P, Bolze AS, Liu Z, Vlamakis H, Graham DB, Buhrlage SJ, Xavier RJ, Clardy J. 2022. *Akkermansia muciniphila* phospholipid induces homeostatic immune responses. *Nature* 608:168–173. <https://doi.org/10.1038/s41586-022-04985-7>
  28. Spragge F, Bakkeren E, Jahn MT, B N Araujo E, Pearson CF, Wang X, Pankhurst L, Cunrath O, Foster KR. 2023. Microbiome diversity protects against pathogens by nutrient blocking. *Science* 382:eadj3502. <https://doi.org/10.1126/science.adj3502>
  29. Li BY, Oh J, Young VB, Rao K, Wiens J. 2019. Using machine learning and the electronic health record to predict complicated *Clostridium difficile* infection. *Open Forum Infect Dis* 6:fz186. <https://doi.org/10.1093/ofid/ofz186>
  30. Seekatz AM, Bassis CM, Fogg L, Moore NM, Rhee Y, Lolans K, Weinstein RA, Lin MY, Young VB, Hayden MK, Centers for Disease Control and Prevention Epicenters Program. 2018. Gut microbiota and clinical features distinguish colonization with *Klebsiella pneumoniae* carbapenemase-producing *Klebsiella pneumoniae* at the time of admission to a long-term acute care hospital. *Open Forum Infect Dis* 5:ofy190. <https://doi.org/10.1093/ofid/ofy190>
  31. Cardile S, Del Chierico F, Candusso M, Reddel S, Bernaschi P, Pietrobattista A, Spada M, Torre G, Putignani L. 2021. Impact of two antibiotic therapies on clinical outcome and gut microbiota profile in liver transplant paediatric candidates colonized by carbapenem-resistant *Klebsiella pneumoniae* CR-KP. *Front Cell Infect Microbiol* 11:730904. <https://doi.org/10.3389/fcimb.2021.730904>
  32. Del Chierico F, Cardile S, Pietrobattista A, Liccardo D, Russo A, Candusso M, Basso MS, Grimaldi C, Pansani L, Bernaschi P, Torre G, Putignani L. 2018. Liver transplantation and gut microbiota profiling in a child colonized by a multi-drug resistant *Klebsiella pneumoniae*: a new approach to move from antibiotic to "Eubiotic" control of microbial resistance. *Int J Mol Sci* 19:1280. <https://doi.org/10.3390/ijms19051280>
  33. Armour CR, Topçuoğlu BD, Garretto A, Schloss PD. 2022. A goldilocks principle for the gut microbiome: taxonomic resolution matters for microbiome-based classification of colorectal cancer. *mBio* 13:e0316121. <https://doi.org/10.1128/mbio.03161-21>
  34. Schloss PD. 2021. Amplicon sequence variants artificially split bacterial genomes into separate clusters. *mSphere* 6:e0019121. <https://doi.org/10.1128/mSphere.00191-21>
  35. Callahan BJ, McMurdie PJ, Holmes SP. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J* 11:2639–2643. <https://doi.org/10.1038/ismej.2017.119>
  36. Martinson JNV, Pinkham NV, Peters GW, Cho H, Heng J, Rauch M, Broadaway SC, Walk ST. 2019. Rethinking gut microbiome residency and the *Enterobacteriaceae* in healthy human adults. *ISME J* 13:2306–2318. <https://doi.org/10.1038/s41396-019-0435-7>
  37. Kang JTL, Teo JY, Bertrand D, Ng A, Ravikrishnan A, Yong M, Ng OT, Marimuthu K, Chen SL, Chng KR, Gan YH, Nagarajan N. 2022. Long-term ecological and evolutionary dynamics in the gut microbiomes of carbapenemase-producing *Enterobacteriaceae* colonized subjects. *Nat Microbiol* 7:1516–1524. <https://doi.org/10.1038/s41564-022-01221-w>
  38. Lam MMC, Wick RR, Watts SC, Cerdeira LT, Wyres KL, Holt KE. 2021. A genomic surveillance framework and genotyping tool for *Klebsiella pneumoniae* and its related species complex. *Nat Commun* 12:4188. <https://doi.org/10.1038/s41467-021-24448-3>
  39. Raffelsberger N, Hetland MAK, Svendsen K, Småbrekke L, Löhr IH, Andreassen LLE, Brisse S, Holt KE, Sundsfjord A, Samuelsen Ø, Gravningen K. 2021. Gastrointestinal carriage of *Klebsiella pneumoniae* in a general adult population: a cross-sectional study of risk factors and bacterial genomic diversity. *Gut Microbes* 13:1939599. <https://doi.org/10.1080/19490976.2021.1939599>
  40. Martin RM, Cao J, Wu W, Zhao L, Manthei DM, Pirani A, Snitkin E, Malani PN, Rao K, Bachman MA. 2018. Identification of pathogenicity-associated loci in *Klebsiella pneumoniae* from hospitalized patients. *mSystems* 3:e00015-18. <https://doi.org/10.1128/mSystems.00015-18>
  41. Seekatz AM, Theriot CM, Molloy CT, Wozniak KL, Bergin IL, Young VB. 2015. Fecal microbiota transplantation eliminates *Clostridium difficile* in a murine model of relapsing disease. *Infect Immun* 83:3838–3846. <https://doi.org/10.1128/IAI.00459-15>
  42. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75:7537–7541. <https://doi.org/10.1128/AEM.01541-09>
  43. Schloss PD. 2009. A high-throughput DNA sequence aligner for microbial ecology studies. *PLoS One* 4:e8230. <https://doi.org/10.1371/journal.pone.0008230>
  44. Westcott SL, Schloss PD. 2017. OptiClust, an improved method for assigning amplicon-based sequence data to operational taxonomic units. *mSphere* 2:e00073-17. <https://doi.org/10.1128/mSphereDirect.00073-17>
  45. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. 2014. Ribosomal database project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res* 42:D633–D642. <https://doi.org/10.1093/nar/gkt1244>
  46. Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 73:5261–5267. <https://doi.org/10.1128/AEM.00062-07>
  47. Oksanen SG, Blanchet F, Kindt R, Legendre P, Minchin P, O'Hara R, Solymos P, Stevens M, Szoeck E, Wagner H, et al. 2022. vegan: community ecology package. V R package version 2.6-4. <https://CRAN.R-project.org/package=vegan>.
  48. Topçuoğlu BD, Lapp Z, Sovacool KL, Snitkin E, Wiens J, Schloss PD. 2021. mikropml: user-friendly R package for supervised machine learning pipelines. *J Open Source Softw* 6:3073. <https://doi.org/10.21105/joss.03073>
  49. Peschel S, Müller CL, von Mutius E, Boulesteix A-L, Depner M. 2020. NetCoMi: network construction and comparison for microbiome data in R. *Bioinformatics*. <https://doi.org/10.1101/2020.07.15.195248>
  50. Friedman J, Alm EJ. 2012. Inferring correlation networks from genomic survey data. *PLoS Comput Biol* 8:e1002687. <https://doi.org/10.1371/journal.pcbi.1002687>
  51. Wickham H. 2016. ggplot2: elegant graphics for data analysis. Springer-Verlag New York, Cham.